A basic way to measure accuracy is the **confusion matrix**:

y	$\hat{y}$	
1	1	True positive
1	-1	False negative
-1	1	False positive
-1	-1	True negative

Synonyms for TP: **recall**, **sensitivity**

Synonyms for TN: **specificity**

Positive predictive value = precision = $\frac{TP}{TP+FP}$

Negative predictive value = $\frac{TN}{TN+FN}$

Misclassification rate = $\frac{FP+FN}{n}$

But misclassification rate is bad for unbalanced classes (e.g. not helpful to always predict that a patient does not have a rare form of cancer). Better measure: **F-score**

$$F = \frac{2(PPV + TPR)}{PPV + TPR}$$

Also better: ROC curve

Proof of why using validation sets will produce the optimal model

Notation

$\{f_\mu\}_{\mu \in \mathcal{M}}$	Models under consideration
$\lvert\mathcal{M}\rvert$	Cardinality of model class (e.g. how many parameters)
$\hat{R}_m()$	misclassification error of model on validation set of size m
$R()$	expected misclassification error of model
$f_{\hat{\mu}}$	model achieving lowest validation error

What we want to show: with probability of at least $1 - \delta$,

$$R(f_{\hat{\mu}}) \leq \min_{\mu \in \mathcal{M}} R(f_\mu) + \sqrt{\frac{2}{m} \log\left(\frac{2|\mathcal{M}|}{\delta}\right)}$$

i.e. as m (the size of validation set) grows, the expected error of the model achieving the lowest validation error attains that of the theoretically optimum model.

How we show it:

Start with the difference in terms of expected risk between the model achieving lowest error on the validation set (henceforth "empirically best model")and the model achieving lowest expected error (henceforth "theoretically best model").

$$R(f_{\hat{\mu}}) - \min_{\mu \in \mathcal{M}} R(f_\mu)$$

Add in canceling terms reflecting the empirical error on validation set of size m of the empirically best model.

$$= R(f_{\hat{\mu}}) - \hat{R}_m(f_{\hat{\mu}}) + \hat{R}_m(f_{\hat{\mu}}) - \min_{\mu \in \mathcal{M}} R(f_\mu)$$

Because we have defined $f_{\hat{\mu}}$ to be the empirically best model, the difference between its empirical risk and the expected risk of the theoretically best model must be less than or equal to the largest possible empirical vs. theoretical risk gap (i.e., the model μ producing the greatest difference between its empirical and expected risk).

$$\leq R(f_{\hat{\mu}}) - \hat{R}_m(f_{\hat{\mu}}) + \max_{\mu \in \mathcal{M}} (\hat{R}_m(f_\mu) - R(f_\mu))$$

And again by definition, the empirical vs. theoretical risk gap of the empirically best model must be less than the empirical vs. theoretical risk gap of the model with the largest such gap.

$$\leq 2 \max_{\mu \in \mathcal{M}} |\hat{R}_m(f_\mu) - R(f_\mu)|$$

Now, consider the probability of that largest possible empirical vs. expected error gap being greater than some arbitrary amount:

$$P(\max_{\mu \in \mathcal{M}} |\hat{R}_m(f_\mu) - R(f_\mu)| > \epsilon)$$

There may be multiple possible models producing that event:

$$= P\left(\bigcup_{\mu \in \mathcal{M}} \{|\hat{R}_m(f_\mu) - R(f_\mu)| > \epsilon)\} \right)$$

Boole's inequality states that $P\left(\cup_i A_i\right) \leq \sum_i P(A_i)$. Thus, the previous inequality must be less than or equal to its summed version

$$\leq \sum_{\mu \in \mathcal{M}} P(|\hat{R}_m(f_\mu) - R(f_\mu)| > \epsilon)$$

Hoeffding's inequality is $P(|\bar{S}_m - E(\bar{S}_m| \geq \epsilon) \leq 2\exp(-2m\epsilon^2/(b-a)^2)$. We are using Bernoulli variables (error calculating using indicator function for whether prediction is correct or not) so $a - b = 1$. Thus:

$$P(|\hat{R}_m(f_\mu) - R(f_\mu)| > \epsilon) \leq 2\exp(-2m\epsilon^2)$$

We plug this back into the previous summation:

$$\sum_{\mu \in \mathcal{M}} P(|\hat{R}_m(f_\mu) - R(f_\mu)| > \epsilon) \leq \sum_{\mu \in \mathcal{M}} 2\exp(-2m\epsilon^2)$$

$$= 2|\mathcal{M}|\exp(-2m\epsilon^2)$$

Use

$$\epsilon = \sqrt{\frac{\log\left(\frac{2|\mathcal{M}|}{\delta}\right)}{2m}}$$

Then, we plug this value into our previous expression

$$2|\mathcal{M}|\exp(-2m\epsilon^2)$$

$$= 2|\mathcal{M}|\exp\left(\frac{-2m\log\left(\frac{2|\mathcal{M}|}{\delta}\right)}{2m}\right)$$

$$= \delta$$

Going all the way back to the start, we now can say

Mallow's CP
Is a thing, but don't feel compelled to take detailed notes on it